

Pragmatic rationalizability

Gerhard Jäger

Gerhard.Jaeger@uni-bielefeld.de

October 2, 2008

Sinn und Bedeutung 13, Stuttgart

*ongoing joint work with Christian Ebert and Roland Mühlenbernd,
SFB 673 "Alignment in Communication", Bielefeld*



Structure of the talk

- Signaling games
- Literal meaning and rationality
- Justification of pragmatic decisions
- Belief revision and forward induction
- Conclusion



Signaling games

- sequential game:
 - 1 **nature** chooses a world w
 - out of a pool of possible worlds W
 - according to a certain probability distribution P
 - 2 nature shows w to sender **S**
 - 3 S chooses a message m out of a set of possible signals M
 - 4 S transmits m to the receiver **R**
 - 5 R chooses an action a , based on the sent message.
- Both S and R have preferences regarding R's action, depending on w .
- S might also have preferences regarding the choice of m (to minimize signaling costs).



Tea or coffee?

An example

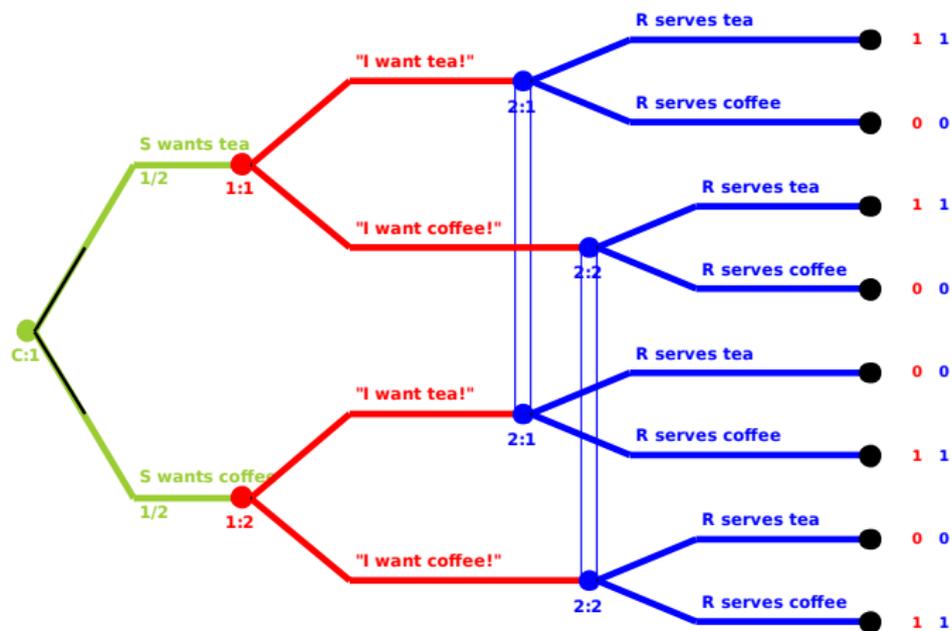
- Sally either prefers tea (w_1) or coffee (w_2).
- Robin either serves tea (a_1) or coffee (a_2).
- Sally can send either of two messages:
 - m_1 : *I prefer tea.*
 - m_2 : *I prefer coffee.*
- Both messages are costless.

	a_1	a_2
w_1	1, 1	0, 0
w_2	0, 0	1, 1

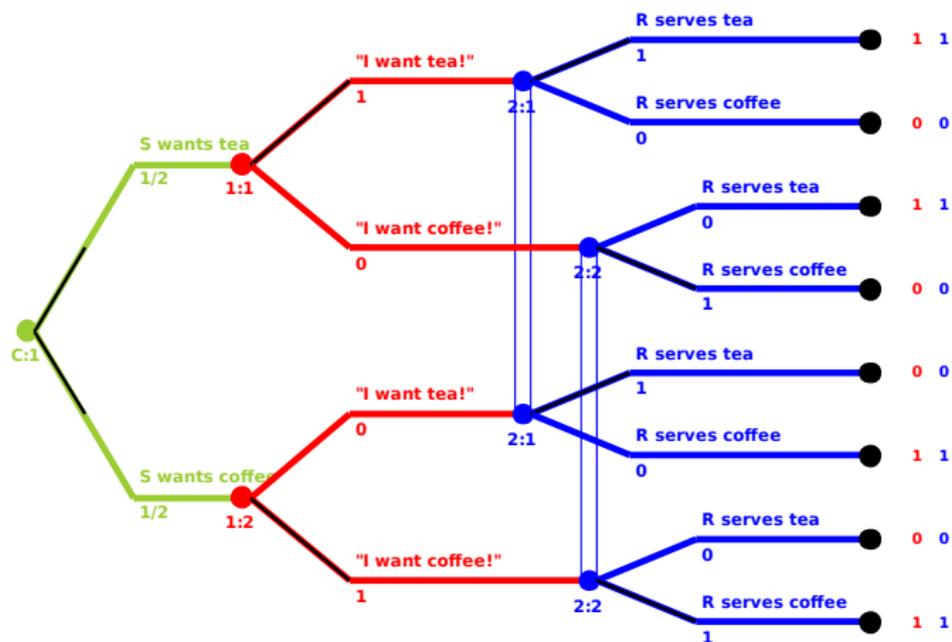
Table: utility matrix



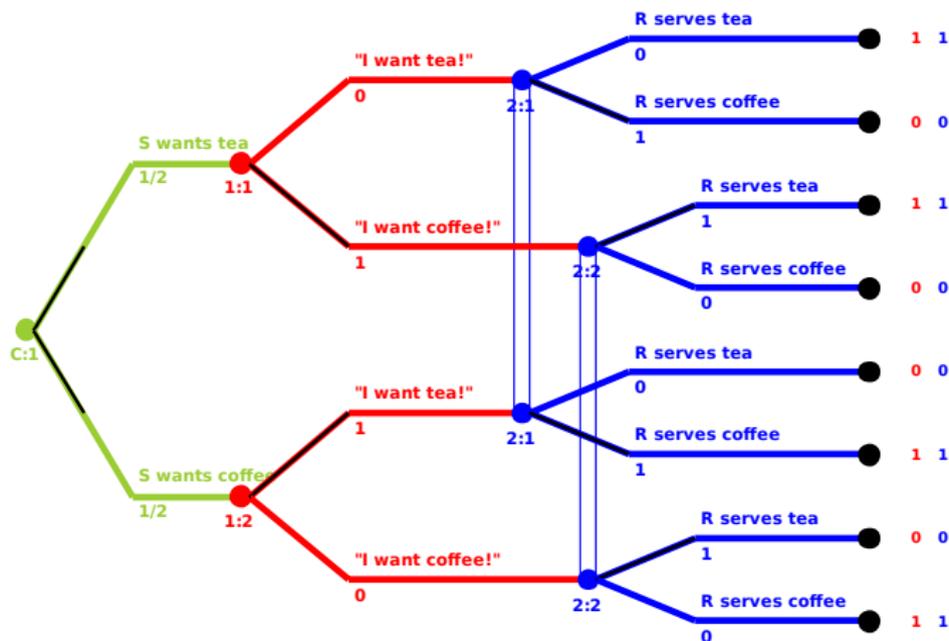
Extensive form



Extensive form



Extensive form



A coordination problem

- two strict Nash equilibria
 - S always says the truth and R always believes her.
 - S always says the opposite of the truth and R interprets everything ironically.
- Both equilibria are equally rational.
- Still, first equilibrium is more reasonable because it employs exogenous meanings of messages for equilibrium selection.
- Criterion for equilibrium selection:

Always say the truth, and always believe what you are told!

- What happens if it is not always rational to be honest/credulous?



Partially aligned interests

Rabin's (1990) example

- In w_1 and w_2 , S and R have identical interests.
- In w_3 , S would prefer R to believe in w_2 .
- The propositions $\{w_1\}$ and $\{w_2, w_3\}$ are *credible*.
- The propositions $\{w_2\}$ and $\{w_3\}$ are *not credible*.

	a_1	a_2	a_3
w_1	10, 10	0, 0	0, 0
w_2	0, 0	10, 10	5, 7
w_3	0, 0	10, 0	5, 7

Table: Partially aligned interests



Partially aligned interests

Rabin's (1990) example

- Suppose there are three messages:
 - m_1 : We are in w_1 .
 - m_2 : We are in w_2 .
 - m_3 : We are in w_3 .
- reasonable S will send m_1 if and only if w_1
- reasonable R will react to m_1 with a_1
- nothing else can be inferred

	a_1	a_2	a_3
w_1	10, 10	0, 0	0, 0
w_2	0, 0	10, 10	5, 7
w_3	0, 0	10, 0	5, 7

Table: Partially aligned interests



**Always say the truth,
and always believe what you are told,
unless you have reasons to do otherwise!**

But what does this mean?



Justification of decisions

- decisions must be **justifiable**
- two kinds of justification:

I use/interpret a message the way I do because:

- *this is what the literal meaning of the message dictates,*
or
- *because this is the best I can do, given my justifiable belief*
about the decisions of the other player.



Sally's belief states

- (first order) belief of the sender:
 - function ρ from messages to probability distribution of actions
 - $\rho(a|m)$: S's subjective probability that R performs action a if S sends message m



Robin's belief states

- (first order) belief of the receiver has two components:
 - function σ_1 from worlds to probability distributions over messages
 - $\sigma_1(m|w)$: R's subjective probability that S sends messages m if she is in world w
 - function σ_2 from messages to probability distributions over worlds
 - $\sigma_2(w|m)$: R's posterior probability that w is the case after observing message m



Bayesian update

- σ_1 and σ_2 are connected via **Bayes' rule**:

$$\sigma_2(w|m) = \frac{\sigma_1(m|w)p^*(w)}{\sum_{w' \in W} \sigma_1(m|w')p^*(w')}$$

provided $\max_{w' \in W} \sigma_1(m|w') > 0$

p^* ... R's prior belief about W



Justification and best responses

Rationality

- A **rational** player will always play a **best response** given their belief state.
- “Best” means: “maximizing the expected utility”

Definition (Best response to beliefs)

$$BR_r(\sigma_2) = \{r \in \mathcal{R} \mid \forall m. r(m) \in \arg \max_{a \in A} \sum_{w \in W} \sigma_2(w|m) u_r(w, m, a)\}$$

$$BR_s(\rho) = \{s \in \mathcal{S} \mid \forall w. s(w) \in \arg \max_{m \in M} \sum_{a \in A} \rho(a|m) u_s(w, m, a)\}$$



Justification and best responses

The outside observer

- Suppose all we know is that Sally believes Robin to play a strategy from some set \mathbf{R} .
- If Sally is rational, she will play a best response to some probability distribution over \mathbf{R} — and vice versa.

Definition (Best response to a set of strategies)

Let $\mathbf{S} \subseteq \mathcal{S}$ and $\mathbf{R} \subseteq \mathcal{R}$ be sets of strategies.

$$BR_r(\mathbf{S}) = \{r \in \mathcal{R} \mid \exists P \in \Delta(\mathbf{S}) \exists \sigma_1 (\forall m, w. \sigma_1(m|w) = \sum_{s \in \mathbf{S}: s(w)=m} P(s)) \\ r \in BR_r(\sigma_2)\}$$

$$BR_s(\mathbf{R}) = \{s \in \mathcal{S} \mid \exists P \in \Delta(\mathbf{R}) \exists \rho (\forall a, m. \rho(a|m) = \sum_{r \in \mathbf{R}: r(m)=a} P(r)) : \\ s \in BR_s(\rho)\}$$

- If Robin is credulous, he will simply believe the literal meaning of each message.
- Formally:
 - Each message m has a literal meaning $\|m\| \subseteq W$.
 - Credulous receiver strategies R_0 : reaction to each message m is an optimal decision after Bayesian update with $\|m\|$:

$$\begin{aligned}\forall w, m. \sigma_2^*(w|m) &= p^*(w|\|m\|) \\ R_0 &= BR_r(\sigma_2^*)\end{aligned}$$



Justification and best responses

- The credulous strategies are justifiable.
- Every best response to a set of credulous strategies is justifiable.
- Nothing else is justifiable.

Definition (Justified strategies: cumulative definition)

$$\begin{aligned}S_0 &= BR_s(R_0) \\J_0 &= (R_0, S_0) \\J_{n+1} &= (S_n \cup BR_s(R_n), R_n \cup BR_r(S_n)) \\J &= \left(\bigcup_n R_n, \bigcup_n S_n\right)\end{aligned}$$



Formal rendering of the maxim given above:

Pragmatic rationalizability:
Play only justifiable strategies!

(cf. Franke 2008a,b, Jäger 2008 for related approaches to pragmatics using iterated best response computation)



Tea and coffee again

Utility matrix

	a_1	a_2
w_1	1, 1	0, 0
w_2	0, 0	1, 1

literal meanings

$$\begin{aligned}\|m_1\| &= \{w_1\} \\ \|m_2\| &= \{w_2\}\end{aligned}$$

$$\begin{aligned}R_0 &= \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2 \end{array} \right] \right\} & S_0 &= \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_2 \end{array} \right] \right\} \\ R_1 &= R_0 & S_1 &= S_0\end{aligned}$$

$$J = (R_0, S_0)$$



Rabin's example

Utility matrix

	a_1	a_2	a_3
w_1	10, 10	0, 0	0, 0
w_2	0, 0	10, 10	5, 7
w_3	0, 0	10, 0	5, 7

literal meanings

$$\|m_1\| = \{w_1\}$$

$$\|m_2\| = \{w_2\}$$

$$\|m_3\| = \{w_3\}$$



Rabin's example

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2/a_3 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$R_2 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2/a_3 \\ m_3 \rightarrow a_2/a_3 \end{array} \right] \right\}$$

$$S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_2 \\ w_3 \rightarrow m_2 \end{array} \right] \right\}$$

$$S_1 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_2/m_3 \\ w_3 \rightarrow m_2/m_3 \end{array} \right] \right\}$$

$$J = (R_2, S_1)$$



(1)

- a. Who came to the party?
- b. m_1 : Some boys came to the party.
- c. m_2 : Some but not all boys came to the party.
- d. m_3 : All boys came to the party.

- $w_1 : \exists x Bx \wedge \forall x. Bx \rightarrow Px$
- $w_2 : \exists x. Bx \wedge Px \wedge \exists x. Bx \wedge \neg Px$
- message costs (to be paid by Sally):
 - $c(m_1) = c(m_3) = 0$
 - $c(m_2) \in \mathbb{R}^+$

	a_1	a_2	a_3
w_1	10, 10	0, 0	8, 8
w_2	0, 0	10, 10	8, 8



Q-implicatures

	a_1	a_2	a_3
w_1	10, 10	0, 0	8, 8
w_2	0, 0	10, 10	8, 8

$$\|m_1\| = \{w_1, w_2\}$$

$$\|m_2\| = \{w_2\}$$

$$\|m_3\| = \{w_1\}$$

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_3 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_1 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_2/a_3 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_1 \end{array} \right] \right\}$$

$$S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_3 \\ w_2 \rightarrow m_1/m_2 \end{array} \right] \right\}$$

$$J = (R_1, S_0)$$



I-implicatures

(2)

- a. John's book is good. \rightsquigarrow The book that John is reading or that he has written is good.
- b. a secretary \rightsquigarrow a female secretary
- c. road \rightsquigarrow hard-surfaced road

formally

- $W = \{w_1, w_2\}$
- $p^*(w_1) = \frac{2}{3}, p^*(w_2) = \frac{1}{3}$
- $\|m_1\| = \{w_1\}, \|m_2\| = \{w_2\},$ and $\|m_3\| = \{w_1, w_2\}$
- $c(m_1) = c(m_2) \in \mathbb{R}^+, c(m_3) = 0$

	a_1	a_2	a_3
w_1	24, 24	0, 0	20, 20
w_2	0, 0	24, 24	16, 16



I-implicatures

	a_1	a_2	a_3
w_1	24, 24	0, 0	20, 20
w_2	0, 0	24, 24	16, 16

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_1/a_3 \end{array} \right] \right\}$$

$$\|m_1\| = \{w_1\} \text{ ("hard-surfaced road")}$$

$$\|m_2\| = \{w_2\} \text{ ("soft-surfaced road")}$$

$$\|m_3\| = \{w_1, w_2\} \text{ ("road")}$$

$$S_0 = \left\{ \left[\begin{array}{l|l|l} w_1 \rightarrow m_1 & m_3 & m_3 \\ w_2 \rightarrow m_2 & m_2 & m_3 \end{array} \right] \right\}$$

$$J = (R_1, S_0)$$



Measure terms

Krifka (2002,2007) notes that measure terms can be used in a precise or in a vague way, and that more complex expressions are less likely to be used in a vague way. Here is a schematic analysis:

- w_1 : 100 meter, w_2 : 101 meter
- m_1 : “one hundred meter”
 m_2 : “one hundred and one meter”
 m_3 : “exactly one hundred meter”
- $\|m_1\| = \|m_3\| = \{w_1\}$,
 $\|m_2\| = \{w_2\}$
- $c(m_1) = 0$, $c(m_2) = c(m_3) \in \mathbb{R}^+$
- a_1 : 100, a_2 : 101, a_3 : [100,101]

	a_1	a_2	a_3
w_1	5, 5	0, 0	3, 3
w_2	0, 0	5, 5	3, 3



Measure terms

	a_1	a_2	a_3
w_1	5, 5	0, 0	3, 3
w_2	0, 0	5, 5	3, 3

$$\|m_1\| = \{w_1\}$$

$$\|m_2\| = \{w_2\}$$

$$\|m_3\| = \{w_1\}$$

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_1 \end{array} \right] \right\}$$

$$S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_2 \end{array} \middle| \begin{array}{l} m_1 \\ m_1 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1/a_3 \\ m_2 \rightarrow a_2 \\ m_3 \rightarrow a_1 \end{array} \right] \right\}$$

$$S_1 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_2 \end{array} \middle| \begin{array}{l} m_3 \\ m_2 \end{array} \middle| \begin{array}{l} m_1 \\ m_1 \end{array} \right] \right\}$$

$$J = (R_1, S_1)$$



How a signal acquires a meaning: Forward induction

The game

If you are driving on a highway in Germany, you sometimes see a car on the opposite lane using the flashlight (“Lichthupe”). It is obvious that the other driver wants you to see his signal, given the timing and the fact that there is no other good reason to use the flashlight. What is going on here? People usually infer that there is some kind of danger ahead (a traffic jam, an accident, or a radar control) and slow down. Why does this work if this is not conventionalized?

(It is conventionalized to some degree, but it seems to me that it would also work without prior precedent.)



How a signal acquires a meaning: Forward induction

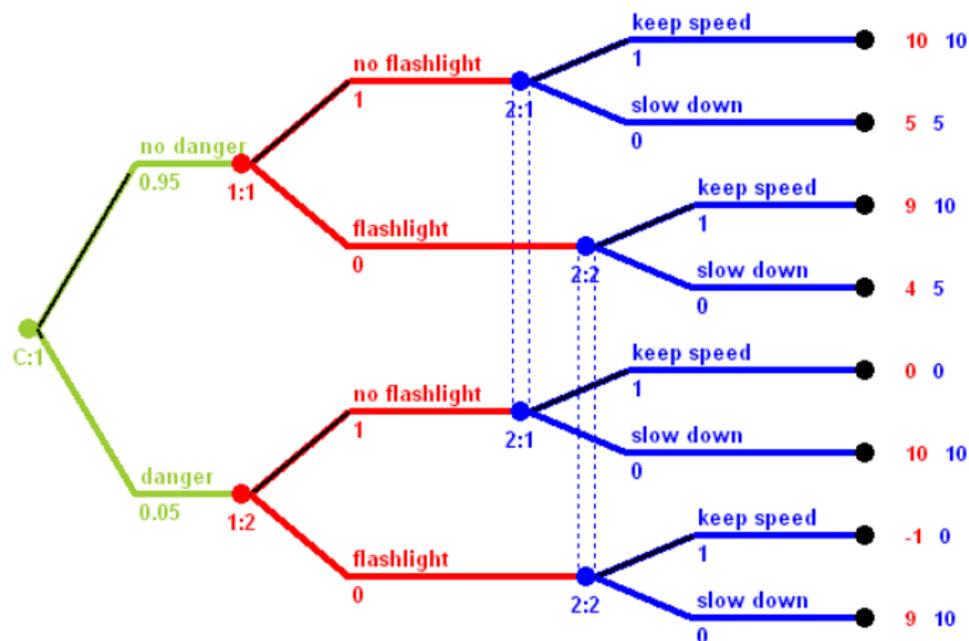
- w_1 : no danger, w_2 : danger
- $p^*(w_1) = \frac{19}{20}, p^*(w_2) = \frac{1}{20}$
- m_1 : no flashlight, m_2 : flashlight
- $c(m_1) = 0, c(m_2) = 1$
- a_1 : keep speed, a_2 : slow down

	a_1	a_2
w_1	10, 10	5, 5
w_2	0, 0	10, 10

Per default, there is no good reason to pay attention to the flashlight, so we expect an equilibrium where nobody uses the flashlight, and nobody would pay attention if somebody used it nevertheless.



Extensive form



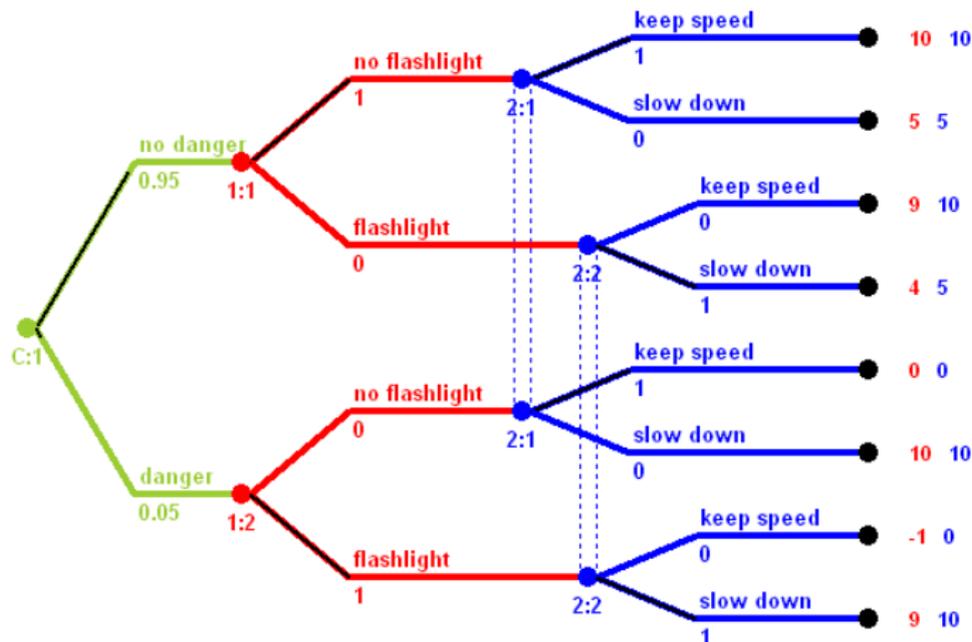
Forward induction

Suppose Robin believes this equilibrium to be common knowledge. If he observes Sally using the flashlight, he might reason: *Sally's behavior has probability 0, so something is wrong with my assumptions. Sally deviates from the equilibrium. The only world where this might be profitable for her is w_2 . Sally is capable to figure out that I will reason this way. So it seems that Sally is in fact in w_2 , and I better slow down because there must be some kind of danger ahead.*

(This kind of reasoning underlies Cho and Kreps' 1987 **Intuitive Criterion**; see also van Rooij 2008.)



Forward induction



What is forward induction?

- In extensive games, players may encounter actions of the other player they considered impossible.
- requires belief revision
- forward induction: even if some previous assumptions must be dropped, belief in rationality of the other player (and her belief in one's own rationality) is preserved if possible
- other belief revision policies may lead to different outcomes



M-implicatures

(4)

- a. m_1 : John stopped the car. \rightsquigarrow John used the foot brake.
- b. m_2 : John made the car stop. \rightsquigarrow John used some other ways to make the car stop.
- c. m_3 : John accelerated the car to sonic speed.

■ w_1/a_1 : foot brake, w_2/a_2 : hand brake, w_3/a_3 : whatever

■ $p^*(w_1) > p^*(w_2) > p^*(w_3)$

■ $\|m_1\| = \|m_2\| = \{w_1, w_2\}, \|m_3\| = \{w_3\}$

■ $c(m_1) = 0, c(m_2), c(m_3) \in \mathbb{R}^+, c(m_2) < c(m_3)$

	a_1	a_2	a_3
w_1	10, 10	1, 1	0, 0
w_2	1, 1	10, 10	0, 0
w_3	0, 0	0, 0	10, 10



M-implicatures

	a_1	a_2	a_3
w_1	10, 10	1, 1	0, 0
w_2	1, 1	10, 10	0, 0
w_3	0, 0	0, 0	10, 10

$$\|m_1\| = \{w_1, w_2\}$$

$$\|m_2\| = \{w_1, w_2\}$$

$$\|m_3\| = \{w_3\}$$

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_1 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_1/??? \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_1 \\ w_3 \rightarrow m_3 \end{array} \right] \right\}$$

What to do with m_2 ???



Belief revision policy

- Forward induction won't do here because it wrongly predicts a_3 as a possible outcome (if only belief in rationality is preserved under belief revision).
- Proposal: first things to revise after unexpected observation:
 - prior probabilities p^*
 - cost function c
- Strong belief (remains constant under belief revision if possible):
 - pragmatic rationalizability



M-implicatures

	a_1	a_2	a_3
w_1	10, 10	1, 1	0, 0
w_2	1, 1	10, 10	0, 0
w_3	0, 0	0, 0	10, 10

$$\|m_1\| = \{w_1, w_2\}$$

$$\|m_2\| = \{w_1, w_2\}$$

$$\|m_3\| = \{w_3\}$$

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_1 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_1 \\ w_3 \rightarrow m_3 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_1/a_2 \\ m_3 \rightarrow a_3 \end{array} \right] \right\}$$

$$S_1 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1 \\ w_2 \rightarrow m_1/m_2 \\ w_3 \rightarrow m_3 \end{array} \right] \right\}$$

$$J = (R_1, S_1)$$



Conversational presuppositions

Simons 2007:

(3)

- a. (Robin:) Are we going to have picnic?
- b. (Sally:) It's raining.

According to Simons, Sally's reply triggers the **presupposition** that *one does not picnic in the rain*—even though there is no conventional presupposition trigger with this effect.



Conversational presuppositions

sketch of an analysis

- $W = \{w_1, w_2, w_3, w_4\}$
- w_1, w_2 : rain, w_3, w_4 : no rain
- w_1, w_3 : rain is bad for picnic,
 w_2, w_4 : rain is not bad for picnic
- $p^*(w_1) < p^*(w_2)$
- m_1 : "It's raining.", m_2 : silence
- $\|m_1\| = \{w_1, w_2\}$,
 $\|m_2\| = \{w_1, w_2, w_3, w_4\}$
- $c(m_1) > c(m_2)$
- a_1 : picnic, a_2 : no picnic

	a_1	a_2
w_1	0, 0	10, 10
w_2	10, 10	0, 0
w_3	10, 10	0, 0
w_4	10, 10	0, 0



Conversational presuppositions

	a_1	a_2
w_1	0, 0	10, 10
w_2	10, 10	0, 0
w_3	10, 10	0, 0
w_4	10, 10	0, 0

$$\|m_1\| = \{w_1, w_2\}$$

$$\|m_2\| = \{w_1, w_2, w_3, w_4\}$$

$$R_0 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1 \\ m_2 \rightarrow a_1 \end{array} \right] \right\} \quad S_0 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_2 \\ w_2 \rightarrow m_2 \\ w_3 \rightarrow m_2 \\ w_4 \rightarrow m_2 \end{array} \right] \right\}$$

$$R_1 = \left\{ \left[\begin{array}{l} m_1 \rightarrow a_1/a_2 \\ m_2 \rightarrow a_1 \end{array} \right] \right\} \quad S_1 = \left\{ \left[\begin{array}{l} w_1 \rightarrow m_1/m_2 \\ w_2 \rightarrow m_2 \\ w_3 \rightarrow m_2 \\ w_4 \rightarrow m_2 \end{array} \right] \right\}$$

$$J = (R_1, S_1)$$



Conversational presuppositions

- Common belief in *conversational presupposition* would be prerequisite to reach equilibrium in (R_0, S_0) .
- If this is not the case, presupposition is accommodated in R_1 , which is employed in S_1 .



Conclusion

- Basic idea: pragmatically rational interlocutors only diverge from literal interpretation if they have good reasons to do so.
- *Good reasons* might be the assumption that the other player employs the literal interpretation, or inferences derived from this and the assumption of rational utility maximization.
- Belief revision policy: first contextual assumptions (about probabilities and preferences) are dropped if needed, semantic knowledge and presumption of rationality are more robust assumptions



Conclusion

- Can be employed: Using an unexpected message amounts to an allusion to a different context (i.e. a different game).
- *I am using this message even though it seems irrational to do so, because there is a different game where it would be rational for you to react to this message in such and such way, and this reaction would be desirable in the present context as well.*
- This inference pattern seems to underly conventional presuppositions, irony, rhetorical questions etc. \rightsquigarrow plenty of topics for further research



- Cho, I.-K. and D. M. Kreps (1987). Signaling games and stable equilibria. *Quarterly Journal of Economics*, **102**(2):179–221.
- Franke, M. (2008a). Interpretation of optimal signals. In K. Apt and R. van Rooij, eds., *New Perspectives on Games and Interaction*, pp. 297–310. Amsterdam University Press, Amsterdam.
- Franke, M. (2008b). Meaning & inference in case of conflict. In K. Balogh, ed., *Proceedings of the 13th ESSLLI Student Session*, pp. 65–74. European Summer School in Logic, Language and Information, Hamburg.
- Jäger, G. (2008). Game theory in semantics and pragmatics. manuscript, University of Bielefeld.
- Krifka, M. (2002). Be brief and vague! and how bidirectional optimality theory allows for verbosity and precision. In D. Restle and D. Zaefferer, eds., *Sounds and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, pp. 439–358. Mouton de Gruyter, Berlin.

- Krifka, M. (2007). Approximate interpretation of number words: A case for strategic communication. In G. Bouma, I. Krämer, and J. Zwarts, eds., *Cognitive foundations of interpretation*, pp. 111–126. Koninklijke Nederlandse Akademie van Wetenschappen, Amsterdam.
- Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, **51**(1):144–170.
- Simons, M. (2007). Presupposition and cooperation. Manuscript, CMU.
- van Rooij, R. (2008). Games and quantity implicatures. To appear in *Journal of Economic Methodology*.

